Original article

Nuclear Medicine Communications

Artificial intelligence in the diagnosis of Parkinson's disease from ioflupane-123 single-photon emission computed tomography dopamine transporter scans using transfer learning

Daniel H. Kim[a], Huub Wit[a] and Mark Thurston[b]

Objective **The objective of this study was to identify the extent to which artificial intelligence could be used in the diagnosis of Parkinson's disease from ioflupane-123 (^{123}I) single-photon emission computed tomography (SPECT) dopamine transporter scans using transfer learning.**

Materials and methods **A data set of 54 normal and 54 abnormal ^{123}I SPECT scans was amplified 44-fold using a process of image augmentation. This resulted in a training set of 2376 normal and 2376 abnormal images. This was used to retrain the top layer of the Inception v3 network. The resulting neural network functioned as a classifier for new ^{123}I SPECT scans as either normal or abnormal. A completely separate set of 45 ^{123}I SPECT scans were used for final testing of the network.**

Results **The area under the receiver-operator curve in final testing was 0.87. This corresponded to a test sensitivity of 96.3%, a specificity of 66.7%, a positive predictive value of 81.3% and a negative predictive value of 92.3%, using an optimum diagnostic threshold.**

Conclusion **This study has provided proof of concept for the use of transfer learning, from convolutional neural networks pretrained on nonmedical images, for the interpretation of ^{123}I SPECT scans. This has been shown to be possible in this study even with a very small sample size. This technique is likely to be applicable to many areas of diagnostic imaging.** ***Nucl Med Commun*** **39:887–893**

Nuclear Medicine Communications 2018, **39**:887–893

Keywords: artificial intelligence, deep learning, image processing, neural networks, Parkinson's disease, transfer learning

[a]The Department of Medical Imaging, The Royal Devon and Exeter NHS Trust, Exeter and [b]The Department of Medical Imaging, Plymouth Hospitals NHS Trust, Plymouth, UK

Correspondence to Daniel H. Kim, FRCR, MBChB, MSc, BSc, The Department of Medical Imaging, The Royal Devon and Exeter NHS Trust, Barrack Road, Exeter EX2 5DW, UK
Tel: +44 779 911 3665; e-mails: dan_kim92@hotmail.com, daniel.kim@nhs.net

Received 8 May 2018 **Revised** 26 June 2018 **Accepted** 16 July 2018

Introduction

Parkinson's disease is an important and common condition with an overall prevalence of 0.3%, increasing to 3% in those aged over 80 years [1]. Signs include motor features such as bradykinesia, rigidity and resting tremor, as well as nonmotor features such as cognitive impairment; disorders of mood, sleep and autonomic function; and sensory symptoms such as pain [2]. Parkinson's disease is part of the Parkinsonian syndrome that is a group of conditions characterized by degeneration of the dopaminergic nigrostriatal pathway and includes the so-called 'Parkinson-plus syndromes', namely progressive supranuclear palsy, multiple system atrophy and corticobasal degeneration [3].

Clinically, it is important to differentiate between Parkinson's syndrome caused by nigrostriatal dopaminergic degeneration and nondegenerative causes of Parkinsonism such as essential tremor or drug-induced Parkinsonism as these conditions follow a different treatment pathway [4]. The ioflupane-123 (^{123}I) single-photon emission computed tomography (SPECT) scan is approved for this purpose and has been shown to have a high specificity [5–10]. The ^{123}I binds to the dopamine transporter and therefore the quantity and distribution of the transporter can be estimated.

In normal ^{123}I SPECT scans there is a characteristic symmetrical uptake in the caudate nuclei and putamen with minimal background activity, giving the morphological appearance of a 'comma' [11]. In Parkinson's disease, the nigrostriatal loss is typically asymmetrical, with preferential loss of uptake in the putamen compared with the caudate nucleus [11–13]. The 'tail of the comma' is therefore often missing in abnormal ^{123}I SPECT scans. Abnormal appearances consistent with Parkinson's disease and Parkinson-plus syndromes also include the symmetrical loss of uptake in both putamen and the complete loss of uptake in the caudate and putamen despite ample background activity [11,14]. A selection of normal and abnormal examples is shown in Fig. 1.

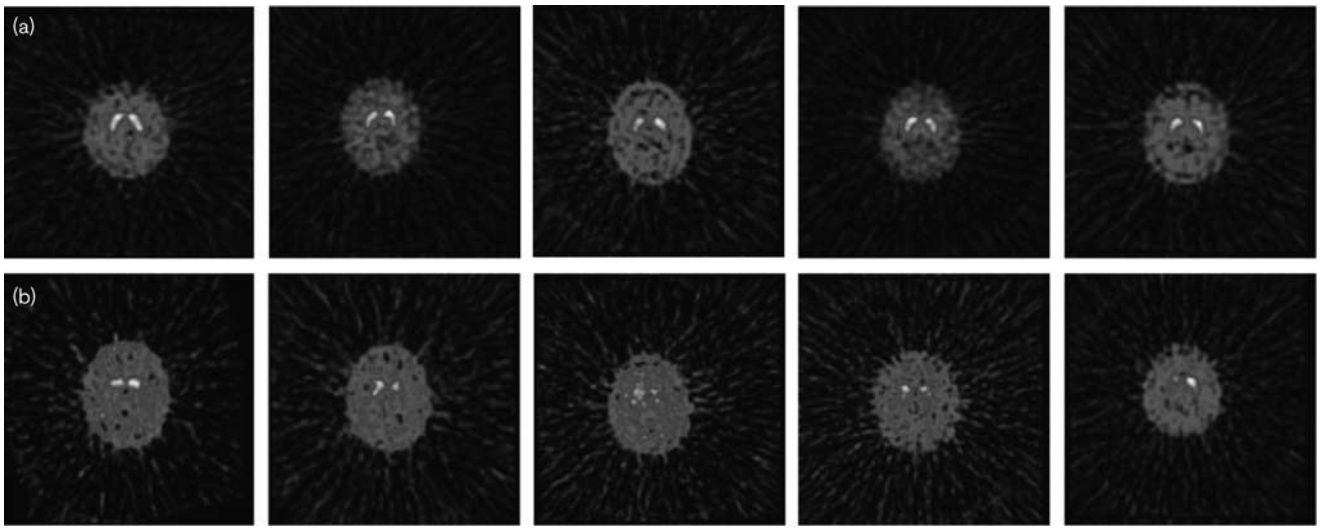
Deep learning

There has been a surge in the use of artificial intelligence (AI) in medical imaging in recent years, with applications

Data were presented previously at the European Conference of Radiology, Vienna, 2018.

DOI: 10.1097/MNM.0000000000000890

Fig. 1



Examples images from ioflupane-123 single-photon emission computed tomography scans as used for training of the neural network. Normal scans are shown in the top row (a) showing the characteristic 'comma' appearance because of symmetrical uptake in the caudate and putamen. Scans consistent with Parkinson's disease and Parkinson's plus syndromes are shown in the bottom row (b) and show either asymmetrical loss of uptake the putamen unilaterally, bilateral loss of uptake in the putamen or near complete loss of uptake in the caudate or putamen despite adequate background activity.

ranging from fracture detection to prediction of longevity [15–17]. Deep learning [18] is a subtype of AI which rose to prominence in 2012 when it was used to win the ImageNet Large Visual Recognition challenge [19]. The technique uses multilayer (deep) artificial neural networks to learn image features associated with a classification. The classification may be a category label such as car or boat, or a diagnosis such as pneumonia or pleural effusion. It is a machine learning method whereby parameters within the model are iteratively updated by a process of error minimization. This requires training data, consisting of a set of prelabelled examples, or ground truth, from which the model can learn from. In this way, the network learns the optimal parameters through exposure to the training data rather than being explicitly programmed.

Training a deep neural network from scratch is computationally demanding and requires large amounts of well-curated data that is often lacking in medical imaging research. However, it is possible to adapt neural networks that have been trained on huge data sets for use in unrelated classification problems. This is known as transfer learning. For example, the Inception v3 network [20] was retrained for the purposes of dermatologist-level classification of skin cancer [21] and for the detection of diabetic retinopathy on retinal fundus photographs [22]. The Inception v3 network was trained on ~1.28 million images for the ImageNet Large Visual Recognition Challenge to recognize over 1000 different image classifications of real-world objects [19,23].

The binary classification of ^{123}I SPECT scans into normal and abnormal is an ideal machine learning challenge. There have been several studies applying techniques such as linear support vector machines and radial basis functions to this problem, some of which achieve excellent accuracy [24–28]. Very high accuracy has also been achieved recently using a three-dimensional convolutional neural network [29]. However, classification of ^{123}I SPECT scans using transfer learning from deep neural networks pretrained on nonmedical images has been underexplored. This study aimed to identify to what extent this was possible and whether this technique could yield clinically useful results with only a small sample size.

Materials and methods

Approval for this study was granted from the UK Health Research Authority. This study was exempt from full ethics review because it was limited to the use of pre-existing, de-identified data.

Image acquisition

^{123}I SPECT scans were obtained from the Royal Devon and Exeter Hospital UK according to a standard protocol. Uptake of unbound radioactive iodine in the thyroid was blocked by administering 170 mg potassium iodate orally, one day before the scan, on the day of the scan and 1 day after the scan. The target activity was 185 MBq ^{123}I (minimum 165 MBq). Imaging commenced at least 3 h after injection and was performed using a dual head gamma camera (Infinia/Optima; GE Healthcare, Chicago,

Illinois, USA). The photopeak window was centred on 159 keV ± 10%, with a 2.5 mm slice thickness and a 128 × 128 matrix (zoom 1.75, 2 × 60 views, 30 s/view, 4° step). Images were reconstructed using filtered back projection with Butterworth prefilter (power 10, critical frequency 0.7).

Data set

^{123}I SPECT scans taking place between February 2015 and July 2017 were included. All scans were reported by a UK consultant radiologist. Scans that showed inconclusive results were excluded. This resulted in 190 scans (118 abnormal and 63 normal). A test set of 45 images (27 abnormal and 18 normal) were randomly selected from the initial sample using a random number generator. This test set was not used in the training process. The size of the training sets was balanced by randomly excluding a further 37 abnormal scans resulting in a training set of 54 normal and 54 abnormal scans. The training sets were balanced to optimize the training of the network.

Data preprocessing and augmentation

The output from a ^{123}I SPECT scan consists of a volume through the basal ganglia. A series of axial slices is then typically produced for clinicians to evaluate. These image series were de-identified and exported in JPEG format. The single slice, most representative of the anatomical location of the basal ganglia, was selected manually for all images (Fig. 2). Because of the small sample size, the training set was augmented using random combinations of rotation, size alteration, shearing and horizontal flip to produce a 44-fold amplification of the training sample size. This resulted in 2376 normal and 2376 abnormal images.

Neural network training

The augmented training data was used to retrain the top layer of the Inception v3 network [20]. The training data were randomly split into training, validation and test sets with an 80 : 10 : 10 ratio. Splitting the data in this way was a trade-off between maximizing the sample size for training the network and minimizing the variance in performance testing and has been used elsewhere [15,23]. Training was performed with a learning rate of 0.01, conducted over 2000 iterations. This was performed using Tensorflow, version 1.0 (Google Brain Team, Apache 2.0 open source license) using the Python 3.5 programming language (Python Software Foundation, Wilmington, Delaware, USA).

Testing the model

The diagnostic performance of the trained model was then evaluated using the test set of 45 images that were not used in the training process. The web-based receiver-operating characteristic analysis tool [30] was used to calculate the area under the receiver-operator curve (AUC). The AUC is used to describe the characteristics of the trained neural network model with higher values indicating better performance and an AUC of 0.5 indicating a test no better than random chance. This was a useful metric in this study for the evaluation of a binary test where the threshold for determining the result can be varied.

Results

A total of 118 (62%) of the 190 ^{123}I SPECT scans performed at the Royal Devon and Exeter hospital showed findings in keeping with Parkinson's or Parkinson's plus syndromes, after removal of inconclusive studies.

A graphical representation of the model training process is illustrated in Fig. 3. After 2000 iterations the error within the network was minimized as showed by a plateau in the loss function (Fig. 3a) with respect to the iteration number. After 2000 iterations the training accuracy plateaued at around 0.95. The difference between this higher training accuracy and the lower final testing accuracy is attributed to overfitting of the network.

The trained neural network produced an output that was a continuous value of between 0 and 1. Values closer to 0 favoured a classification of abnormal whereas values closer to 1 favoured normal. The AUC was 0.87, showing excellent levels of diagnostics test accuracy (Fig. 4).

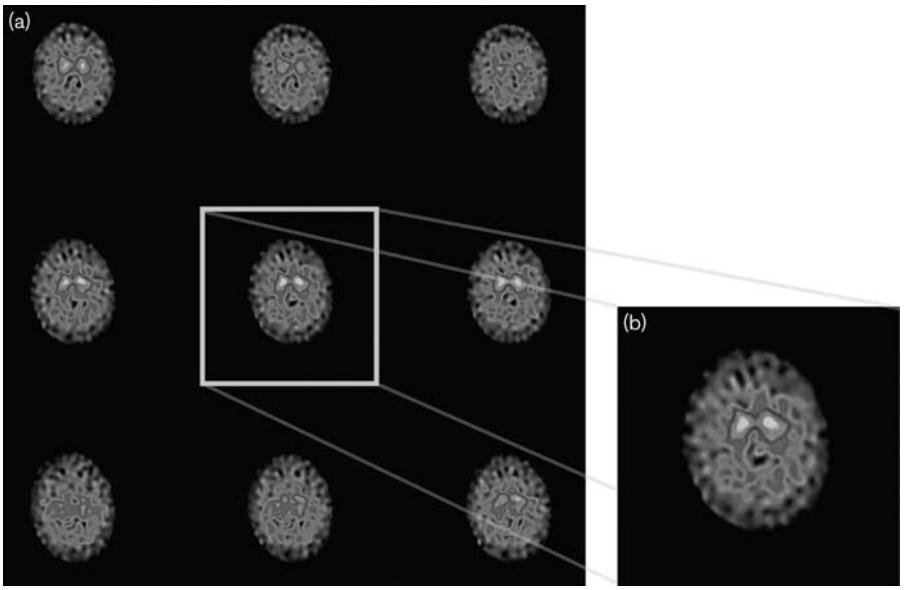
The threshold for classification was optimized through analysis of the ROC. Clinical utility of the model as a diagnostic test could be optimized by setting a score threshold of 0.395 (Fig. 5). This would result in a test sensitivity of 96.3% and specificity of 66.7%. Using the same threshold would produce a positive predictive value of 81.3% and a negative predictive value of 92.3% (Table 1).

Discussion

This study is a proof of concept that high diagnostic test performance can be achieved using deep learning for the analysis of ^{123}I SPECT scans. This is achievable even with a very small sample size by leveraging the power of large pretrained neural networks through the process of transfer learning. This is because features learned in early network layers represent the building blocks of images in general, such as lines, curves and textures, which are translatable to a wide variety of image types [31]. This study provides evidence that these basic image features derived from everyday images such as cats and vehicles are also translatable to medical images from nuclear medicine studies.

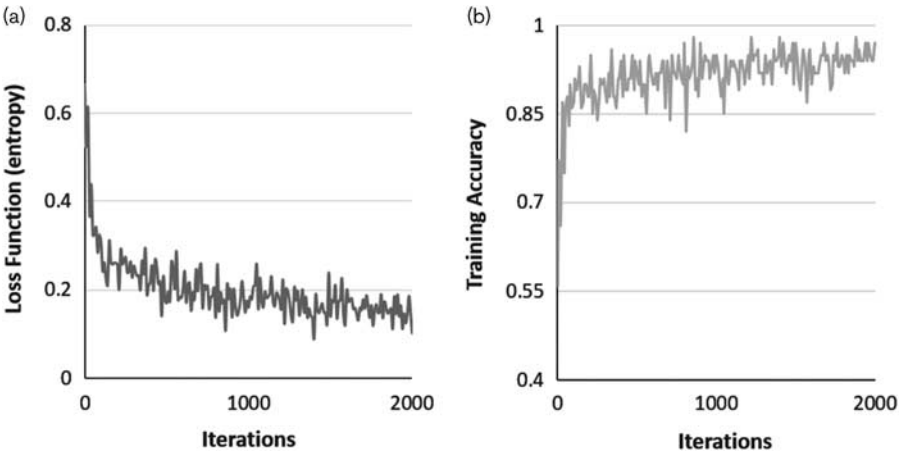
The sample size of 190 images used in the study is extremely low when compared with the numbers typically used in deep learning studies. Moreover, it is orders of magnitude less than the numbers typically used to accurately train a neural network from scratch. For example, a recent study used over 100 000 images to train

Fig. 2



An example of the axial series produced for evaluation by clinicians is shown (a). The white bounding box shows the section most anatomically representative of the basal ganglia. The final image used for training the network is also shown (b).

Fig. 3



Graphical representation of the model training process. The loss function (a) and training accuracy (b) are shown with respect to the iteration number.

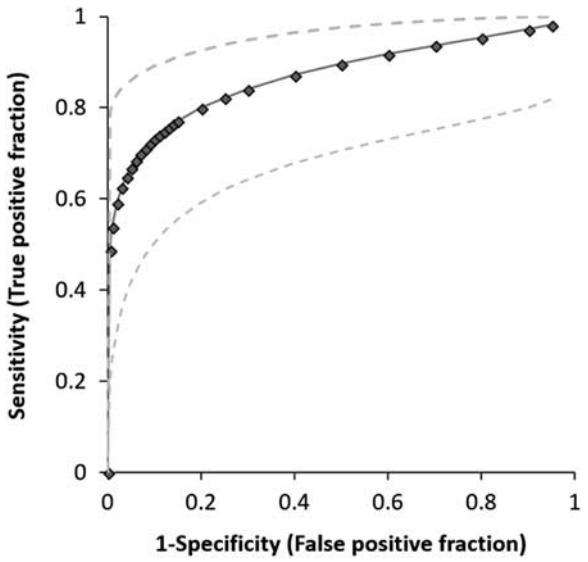
a neural network for the detection of pneumonia on chest radiographs [32]. Large data sets of accurately labelled medical imaging data can be difficult or costly to obtain. The reduction in sample size requirements described in this study is, therefore, promising for the use of transfer learning in other areas of medical imaging where sample size is limited.

Although the levels of accuracy presented here fall slightly short of those published elsewhere [24,29] the findings are an important proof of concept for the use of transfer learning with data augmentation as applied to ^{123}I SPECT scans. This is because the method achieves high levels of accuracy with a much smaller sample size than has previously been required for deep learning in medical imaging. This is important because the availability of large accurately labelled data sets has been one of the largest barriers to clinical implementation of AI technology.

Limitations

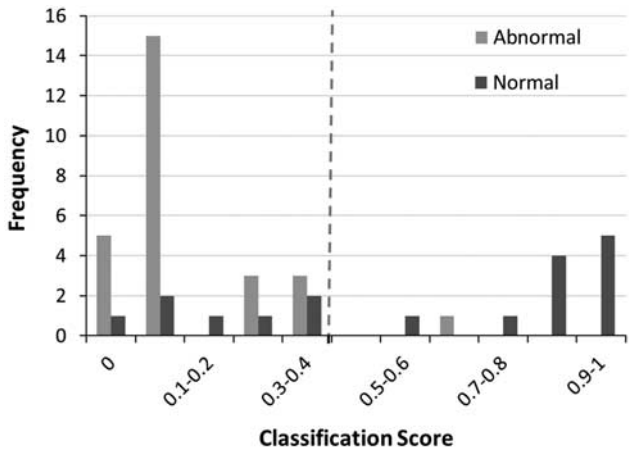
This was not a fully automated diagnostic pipeline. The need to manually select the single image that best

Fig. 4



A fitted receiver-operator curve (solid line with diamond data points) illustrating sensitivity with respect to sensitivity (1 − sensitivity). Dashed lines represent the upper and lower 95% confidence intervals, respectively.

Fig. 5



Histogram illustrating the frequency distribution of classification scores with respect to the ground truth. The network predicts that the image is abnormal when scores are closer to 0 and normal when scores are closer to 1. The true classification (ground truth) is illustrated by the differently shaded bars. The threshold between a normal and abnormal classification is shown by the dashed grey line.

represented the anatomical location of the basal ganglia was a limiting factor. This was a potential source of bias; however, this selection process only required anatomical knowledge rather than expert knowledge of pathology. This step could be automated by using consecutive neural networks in which the first network identified the most relevant image based on the anatomy. The image identified from the first network could then be fed into a separate classification network, such as the network described in this study. A similar methodology using cascaded fully convolutional neural networks has recently been validated elsewhere [33]. Using the entire scan volume rather than a single slice might be a productive area of future research. This would require a suitable pretrained three-dimensional neural network if the transfer learning methodology is to be used.

Table 1 Contingency table illustrating test result with respect to ground truth

	Truth		
Tests	Normal	Abnormal	Total
Normal	12	1	13
Abnormal	6	26	32
Total	18	27	–

The neural network was trained using images from the series of images sent to the picture archiving and communications system from the ^{123}I SPECT scan data. These were preprocessed from a dedicated image viewer so that the pixel intensities were appropriate for diagnostic interpretation. This process is similar to 'windowing' in computed tomography where the Hounsfield units are mapped to the most appropriate pixel values for visualization on a computer screen. This may have introduced a source of bias. For example, the background pixel values may have been susceptible to artificial alterations introduced in the 'windowing' process, especially in scans where the dopamine transporter uptake was reduced. This could be overcome by using the source data and applying normalization from the histogram of pixel values. However, in the process of windowing the scan image, the background pixel values should consistently change proportionately to the pixel values in the basal ganglia. Therefore, the network interpretation of the resulting image is likely to be meaningful regardless of the changes in windowing. This could be further investigated using filter visualization or an activation heat map to indicate the regions of the image most contributing to the classification score, as described elsewhere [34]. This would also reduce the perception of the deep learning technique as a 'black-box' method and allow for more rigorous scrutiny of the underlying mechanisms.

The design of this study meant that it was impossible for the neural network to outperform the radiologist as there was no ground truth superior to the radiologists' opinion. In future studies, a superior ground truth could be established by following up patients over a longer time period to confirm the diagnosis as it emerges over time or by pooling consensus from multiple expert radiologists. In the latter, the performance of any single radiologist could then be compared with the performance of the

neural network. Recent studies using this approach have been able to show superior neural network performance compared with a radiologist in specific pathology identification tasks [32].

The lower specificity showed in this study would almost certainly be improved by increasing the sample size or by adjusting the threshold to reduce false positives. The threshold could be modified to take account of the clinical context depending on the relative importance of disease identification with respect to misdiagnosis.

In this study, the inception V3 network was chosen for retraining; however, there are several other pretrained networks that could be used for this purpose such as ResNet or VGG models [35,36]. Future studies could compare the retraining of multiple different pretrained models as the most effective architecture for nuclear medicine scans of this type and quantity is not yet known. Furthermore, retraining of more than just the final layer of the pretrained network might achieve better performance.

Conclusion

Automated analysis of radiological images with an AUC of 0.87, a sensitivity of 96% and a negative predictive value of 92% is an exciting proof of concept for AI in nuclear medicine. It highlights the use of transfer learning in achieving high performance with a very small sample size. This study also shows that neural networks trained on nonmedical images can learn features that are relevant in the analysis of medical imaging, including in nuclear medicine. This technique could be adopted across many different imaging investigations and may be a useful tool for improving the efficiency and accuracy of clinical radiologists.

Acknowledgements

The authors would like the thank the Royal Devon and Exeter NHS Trust for sponsoring this study and the Research and Development team at the Royal Devon and Exeter NHS Trust for their support.

Conflicts of interest

There are no conflicts of interest.

References

1 Pringsheim T, Jette N, Frolkis A, Steeves TDL. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov Disord* 2014; **29**:1583–1590.
2 Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, *et al*. MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord* 2015; **30**:1591–1601.
3 Seifert KD, Wiener JI. The impact of DaTscan on the diagnosis and management of movement disorders: a retrospective study. *Am J Neurodegener Dis* 2013; **2**:29–34.
4 Pagano G, Niccolini F, Politis M. Imaging in Parkinson's disease. *Clin Med (Lond)* 2016; **16**:371–375.
5 Benamer HTS, Patterson J, Grosset DG, Booij J, De Bruin K, Van Royen E, *et al*. Accurate differentiation of parkinsonism and essential tremor using visual assessment of [^{123}I]-FP-CIT SPECT imaging: The [^{123}I]-FP-CIT study group. *Mov Disord* 2000; **15**:503–510.
6 Benamer HTS, Oertel WH, Patterson J, Hadley DM, Pogarell O, Höffken H, *et al*. Prospective study of persynaptic dopaminergic imaging in patients with mild parkinsonism and tremor disorders: Part 1. baseline and 3-month observations. *Mov Disord* 2003; **18**:977–984.
7 Jennings DL, Seibyl JP, Oakes D, Eberly S, Murphy J, Marek K. (^{123}I) beta-CIT and single-photon emission computed tomographic imaging vs clinical evaluation in Parkinsonian syndrome: unmasking an early diagnosis. *Arch Neurol* 2004; **61**:1224–1229.
8 Stoessl AJ, Lehericy S, Strafella AP. Imaging insights into basal ganglia function, Parkinson's disease, and dystonia. *Lancet* 2014; **384**:532–544.
9 Politis M. Neuroimaging in Parkinson disease: from research setting to clinical practice. *Nat Rev Neurol* 2014; **10**:708–722.
10 Tolosa E, Vander Borght T, Moreno E. Accuracy of DaTSCAN (^{123}I-Ioflupane) SPECT in diagnosis of patients with clinically uncertain parkinsonism: 2-year follow-up of an open-label study. *Mov Disord* 2007; **22**:2346–2351.
11 Booth TC, Nathan M, Waldman AD, Quigley AM, Schapira AH, Buscombe J. The role of functional dopamine-transporter SPECT imaging in parkinsonian syndromes, part 2. *Am J Neuroradiol* 2015; **36**:236–244.
12 Innis RB, Seibyl JP, Scanley BE, Laruelle M, Abi-Dargham A, Wallace E, *et al*. Single photon emission computed tomographic imaging demonstrates loss of striatal dopamine transporters in Parkinson disease. *Proc Natl Acad Sci USA* 1993; **90**:11965–11969.
13 Brucke T, Asenbaum S, Pirker W, Djamshidian S, Wenger S, Wober C, *et al*. Measurement of the dopaminergic degeneration in Parkinson's disease with [^{123}I] beta-CIT and SPECT. Correlation with clinical findings and comparison with multiple system atrophy and progressive supranuclear palsy. *J Neural Transm Suppl* 1997; **50**:9–24.
14 Catafau AM, Tolosa E. Impact of dopamine transporter SPECT using ^{123}I-Ioflupane on diagnosis and management of patients with clinically uncertain Parkinsonian syndromes. *Mov Disord* 2004; **19**:1175–1182.
15 Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol* 2018; **73**:439–445.
16 Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ. Precision Radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework. *Sci Rep* 2017; **7**:1–13.
17 Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol* 2017; **10**:257–273.
18 Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**:436–444.
19 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, *et al*. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; **115**:211–252.
20 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2015. Available at: *http://arxiv.org/abs/1512.00567*. [Accessed 7 May 2018].
21 Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**:115–118. [Accessed 7 May 2018].
22 Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316**:2402–2410.
23 Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, *et al*. Going deeper with convolutions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; New York, NY: CVPR; 2015. pp. 1–9.
24 Augimeri A, Cherubini A, Cascini GL, Galea D, Caligiuri ME, Barbagallo G, *et al*. CADA – computer-aided DaTSCAN analysis. *EJNMMI Phys* 2016; **3**:1.
25 Oliveira FPM, Castelo-Branco M. Computer-aided diagnosis of Parkinson's disease based on [(123)I]FP-CIT SPECT binding potential images, using the voxels-as-features approach and support vector machines. *J Neural Eng* 2015; **12**:26008.
26 Prashanth R, Roy SD, Mandal PK, Ghosh S. High-accuracy classification of Parkinson's disease through shape analysis and surface fitting in ^{123}I-ioflupane SPECT imaging. *IEEE J Biomed Health Inform* 2017; **21**:794–802.
27 Tagare HD, DeLorenzo C, Chelikani S, Saperstein L, Fulbright RK. Voxel-based logistic analysis of PPMI control and Parkinson's disease DaTscans. *Neuroimage* 2017; **152**:299–311.
28 Taylor JC, Fenner JW. Comparison of machine learning and semi-quantification algorithms for (I123)FP-CIT classification: the beginning of the end for semi-quantification? *EJNMMI Phys* 2017; **4**:29.
29 Choi H, Ha S, Im HJ, Paek SH, Lee DS. Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *NeuroImage Clin* 2017; **16**:586–594.

30 Eng J. *ROC analysis: web-based calculator for ROC curves*. Baltimore, MD: Balt John Hopkins Univ.
31 Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? CoRR. 2014. Available at: *http://arxiv.org/abs/1411.1792*. [Accessed 7 May 2018].
32 Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, *et al*. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. 2017. Available at: *http://arxiv.org/abs/1711.05225*. [Accessed 7 May 2018].
33 Christ PF, Ettlinger F, Grün F, Elshaer MEA, Lipková J, Schlecht S, *et al*. Automatic liver and tumor segmentation of {CT} and {MRI} volumes using cascaded fully convolutional neural networks. CoRR. 2017. Available at: *http://arxiv.org/abs/1702.05970*. [Accessed 7 May 2018].
34 Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Networks Learn Syst* 2017; **28**:2660–2673.
35 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015; Available at: *https://arxiv.org/abs/1512.03385*. [Accessed 7 May 2018].
36 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. Available at: *https://arxiv.org/abs/1409.1556*. [Accessed 7 May 2018].